PYTHON ON DATA SCIENCE PROJECT REPORT

EXPLORATORY DATA ANALYSIS

GROUP MEMBERS

❖	E Jaswanth Krishna	: S20200020257
❖	K Balarajaiah	: S20200010085

INTRODUCTION

Heart attack prediction is a critical area of research, as heart disease is one of the leading causes of death worldwide. In this report, we will explore the Heart.csv dataset and perform exploratory data analysis (EDA) to gain insights into the data and build a predictive model for heart attack risk.

By performing EDA on this dataset, we aim to identify any patterns, correlations, or trends in the data that can help us understand the relationship between different variables and the target variable, i.e., the likelihood of having a heart attack. We will explore various visualization techniques such as histograms, scatterplots, and boxplots to gain a deeper understanding of the data.

ABSTRACT

The Heart.csv dataset contains 14 columns and 303 rows of data collected from patients with suspected heart disease. The columns include age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and the target variable indicating the presence of heart disease.

- Age : Age of the patient
- Sex : Sex of the patient
- exang: exercise induced angina (1 = yes; 0 = no)
- ca: number of major vessels (0-3)
- cp : Chest Pain type chest pain type
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic
- trtbps : resting blood pressure (in mm Hg)
- chol : cholestoral in mg/dl fetched via BMI sensor
- fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- rest_ecg : resting electrocardiographic results
    - Value 0: normal
    - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach : maximum heart rate achieved
- target : 0= less chance of heart attack 1= more chance of heart attack

MOTIVATION

The motivation for heart attack prediction is to improve the health outcomes of individuals by identifying those who are at risk of heart disease and providing timely preventive measures and treatments. Heart disease is a leading cause of death worldwide, and early detection and prevention can save lives and reduce healthcare costs. By building accurate predictive models for heart attack risk, we can help healthcare providers make informed decisions, develop personalized treatment plans, and improve patient outcomes. Additionally, this research can contribute to the development of new preventive measures and treatments for heart disease, ultimately improving public health.

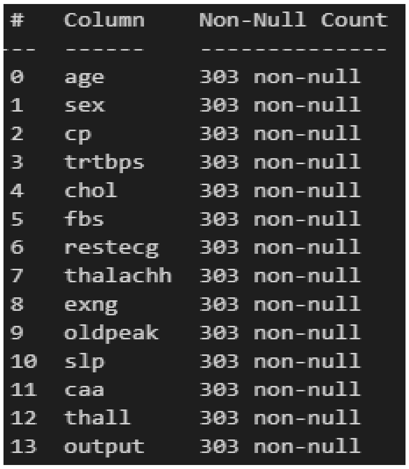
PROBLEM STATEMENT

We want to create a better way to predict the risk of heart attack based on a person's age, medical history, and other factors. This is important because current methods may not be accurate enough to identify those at risk of heart disease, which can be deadly. By using a dataset called Heart.csv, we hope to develop a more accurate and personalized way to predict heart attack risk, which could help prevent heart disease and save lives.

METHODOLOGY

1. DATA PREPROCESSING

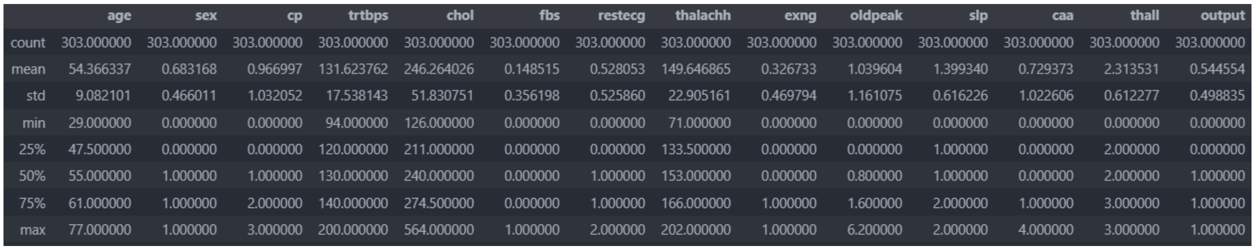
a. Null values: There are no null values in data.



```
#   Column    Non-Null Count
--  ------    --------------
0   age       303 non-null
1   sex       303 non-null
2   cp        303 non-null
3   trtbps    303 non-null
4   chol      303 non-null
5   fbs       303 non-null
6   restecg   303 non-null
7   thalachh  303 non-null
8   exng      303 non-null
9   oldpeak   303 non-null
10  slp       303 non-null
11  caa       303 non-null
12  thall     303 non-null
13  output    303 non-null
```

b. Summary:

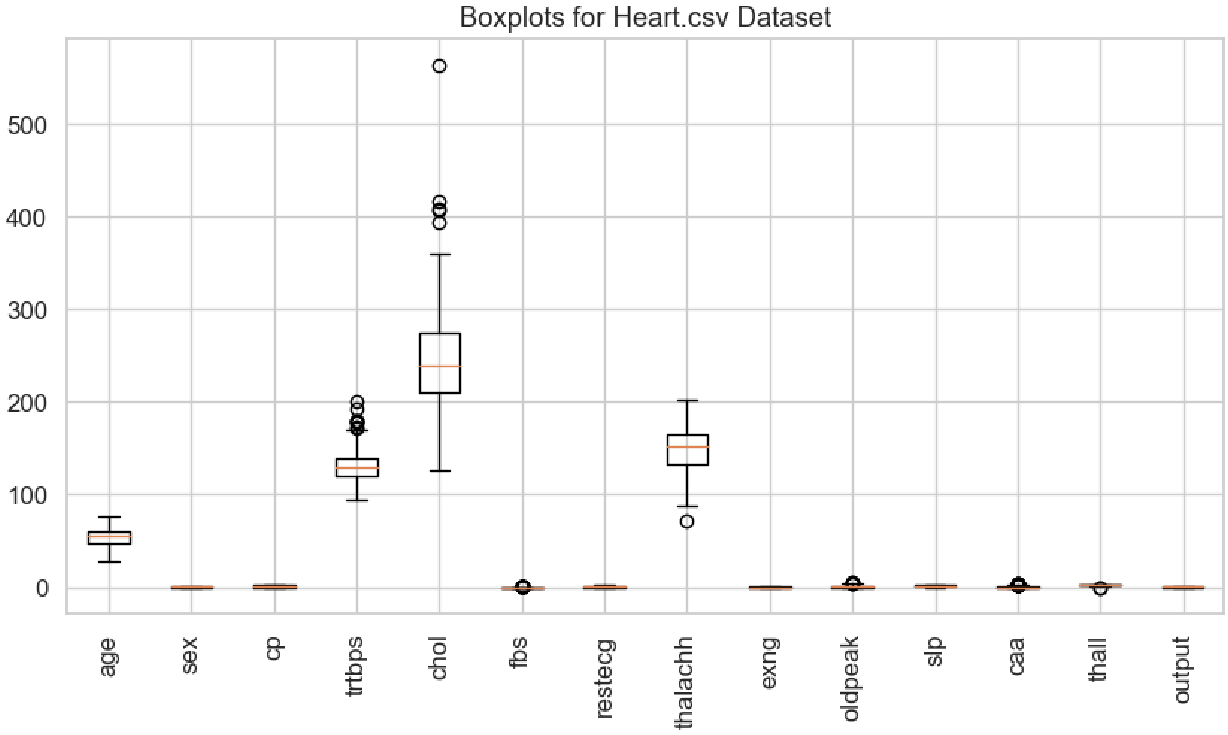
This summary about dataset heart.csv. Here we can find the count,mean,std,min,max,IQR ranges in each row of data.



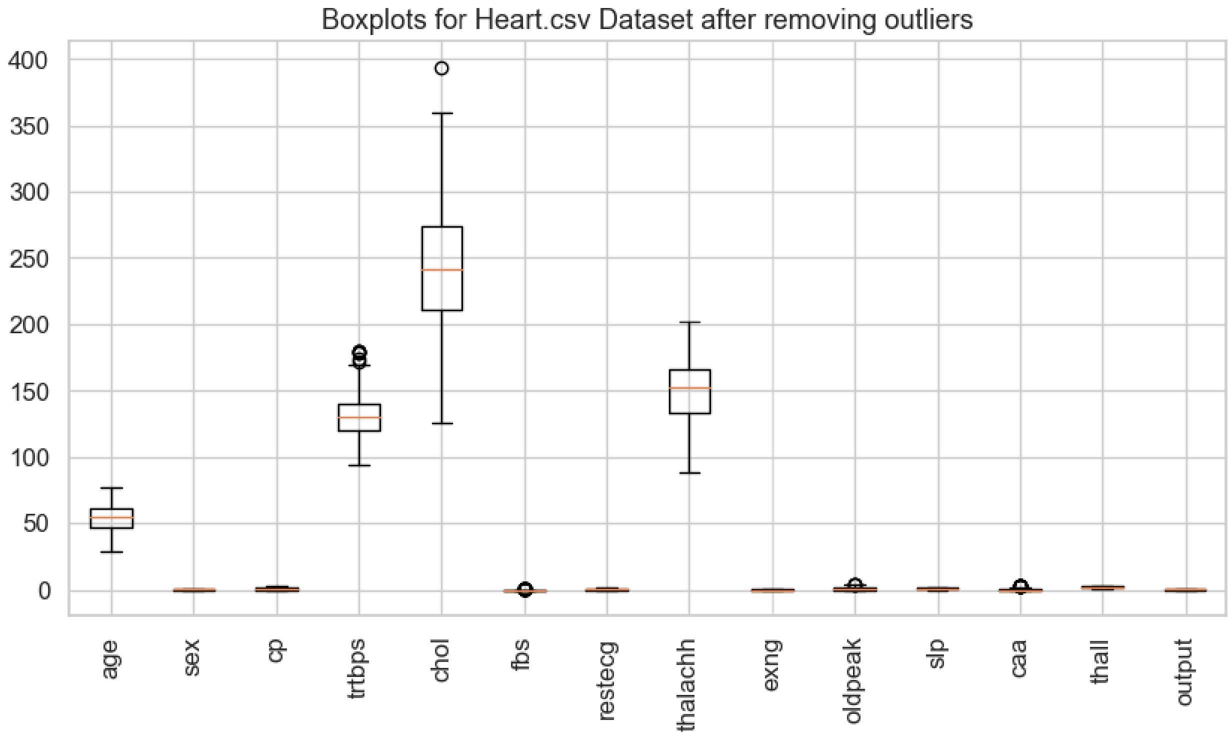
	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

INFLUENTIAL POINTS:

The points which are not in range of IQR are called outliers in x direction and in target direction are called as leverage points commonly known as influential points.



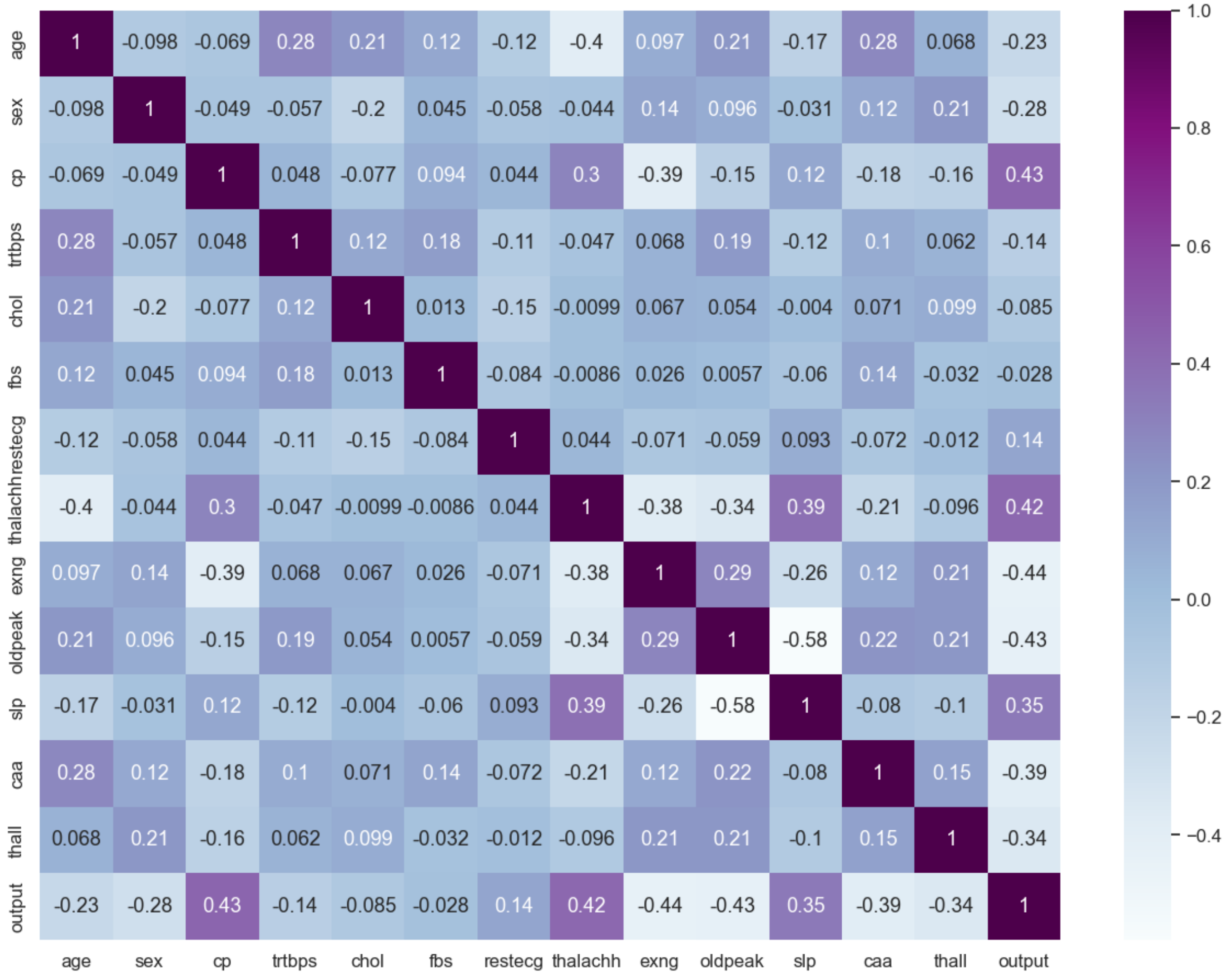
☐ So we have removed outliers by Z-Score values.The rows with z-score greater than 3 are called as outliers.Box plots after outliers removal.



2. EXPLORATORY DATA ANALYSIS(EDA)

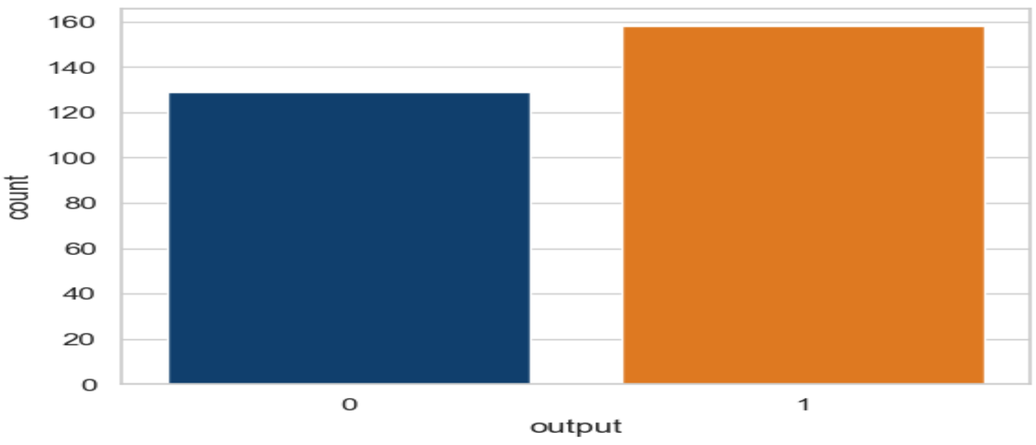
a. HEAT_MAP

Here we see the correlation between all attributes in the dataset. And we can see how they are correlated. If the correlation between them is nearer to 1 are called as that columns are correlated.Here the highest correlation is 0.43 for cp and output column.And the are values in 0.00 terms defines that columns are highly independent by this heat map.The below figure the correlation between each and every attributes

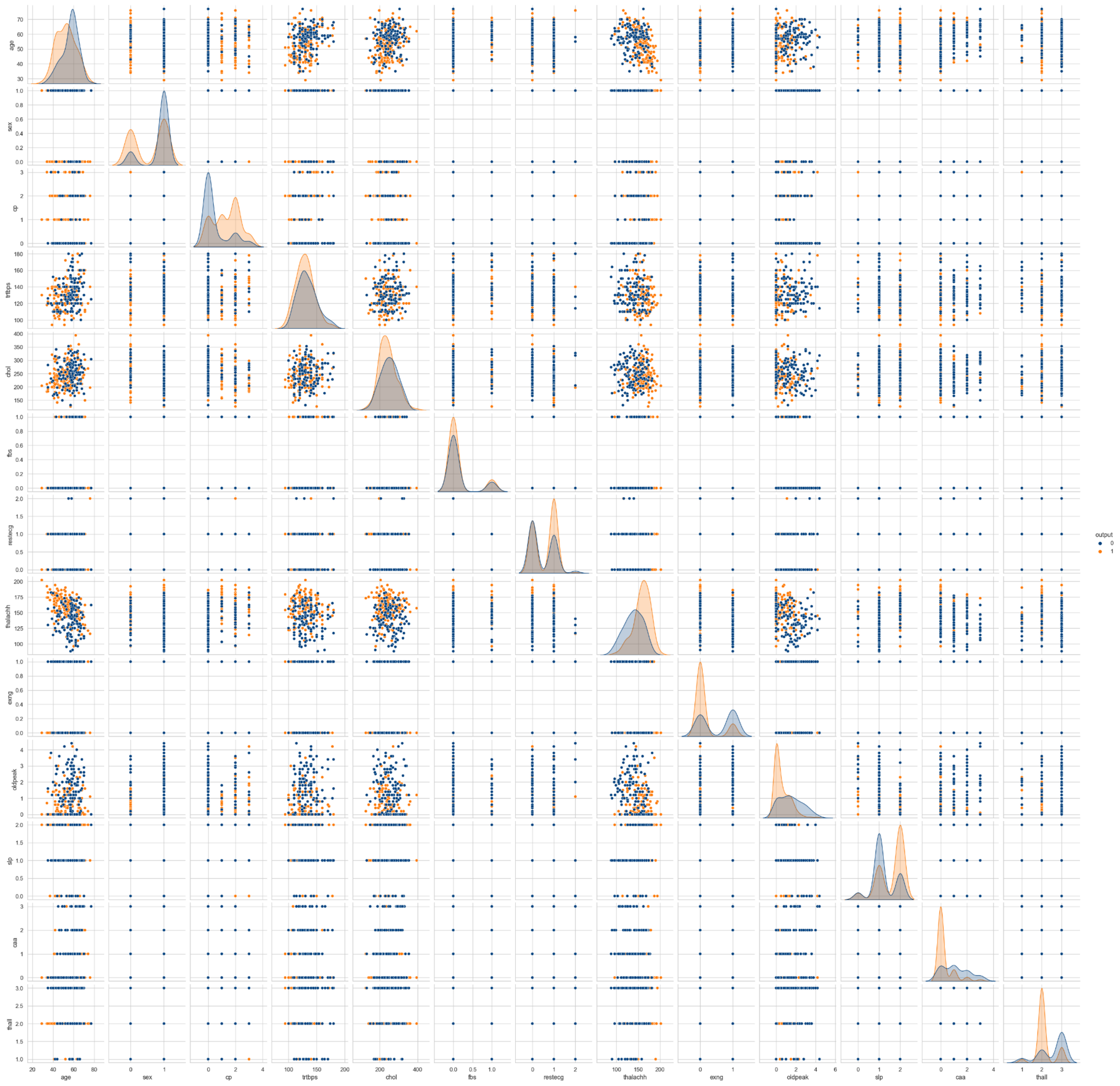


	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
age	1	-0.098	-0.069	0.28	0.21	0.12	-0.12	-0.4	0.097	0.21	-0.17	0.28	0.068	-0.23
sex	-0.098	1	-0.049	-0.057	-0.2	0.045	-0.058	-0.044	0.14	0.096	-0.031	0.12	0.21	-0.28
cp	-0.069	-0.049	1	0.048	-0.077	0.094	0.044	0.3	-0.39	-0.15	0.12	-0.18	-0.16	0.43
trtbps	0.28	-0.057	0.048	1	0.12	0.18	-0.11	-0.047	0.068	0.19	-0.12	0.1	0.062	-0.14
chol	0.21	-0.2	-0.077	0.12	1	0.013	-0.15	-0.0099	0.067	0.054	-0.004	0.071	0.099	-0.085
fbs	0.12	0.045	0.094	0.18	0.013	1	-0.084	-0.0086	0.026	0.0057	-0.06	0.14	-0.032	-0.028
restecg	-0.12	-0.058	0.044	-0.11	-0.15	-0.084	1	0.044	-0.071	-0.059	0.093	-0.072	-0.012	0.14
thalachh	-0.4	-0.044	0.3	-0.047	-0.0099	-0.0086	0.044	1	-0.38	-0.34	0.39	-0.21	-0.096	0.42
exng	0.097	0.14	-0.39	0.068	0.067	0.026	-0.071	-0.38	1	0.29	-0.26	0.12	0.21	-0.44
oldpeak	0.21	0.096	-0.15	0.19	0.054	0.0057	-0.059	-0.34	0.29	1	-0.58	0.22	0.21	-0.43
slp	-0.17	-0.031	0.12	-0.12	-0.004	-0.06	0.093	0.39	-0.26	-0.58	1	-0.08	-0.1	0.35
caa	0.28	0.12	-0.18	0.1	0.071	0.14	-0.072	-0.21	0.12	0.22	-0.08	1	0.15	-0.39
thall	0.068	0.21	-0.16	0.062	0.099	-0.032	-0.012	-0.096	0.21	0.21	-0.1	0.15	1	-0.34
output	-0.23	-0.28	0.43	-0.14	-0.085	-0.028	0.14	0.42	-0.44	-0.43	0.35	-0.39	-0.34	1

□ The following jointplot shows counts of output columns, which means factors lead to heart stroke or not.

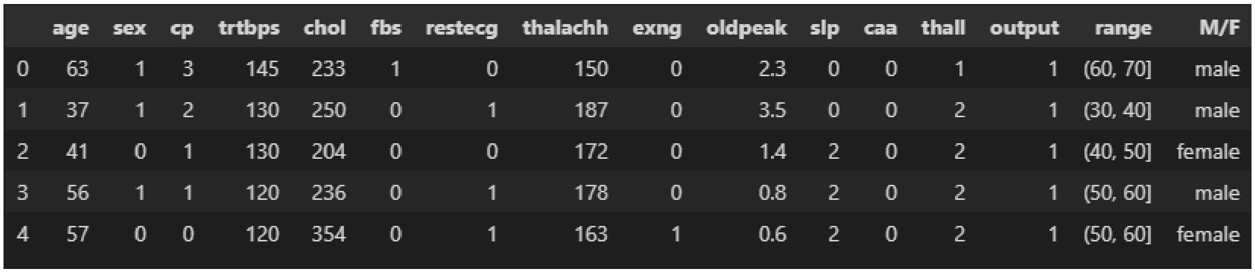


□ Here we can see the pair plot which takes only numerical attributes and gives interrelation between any two attributes.



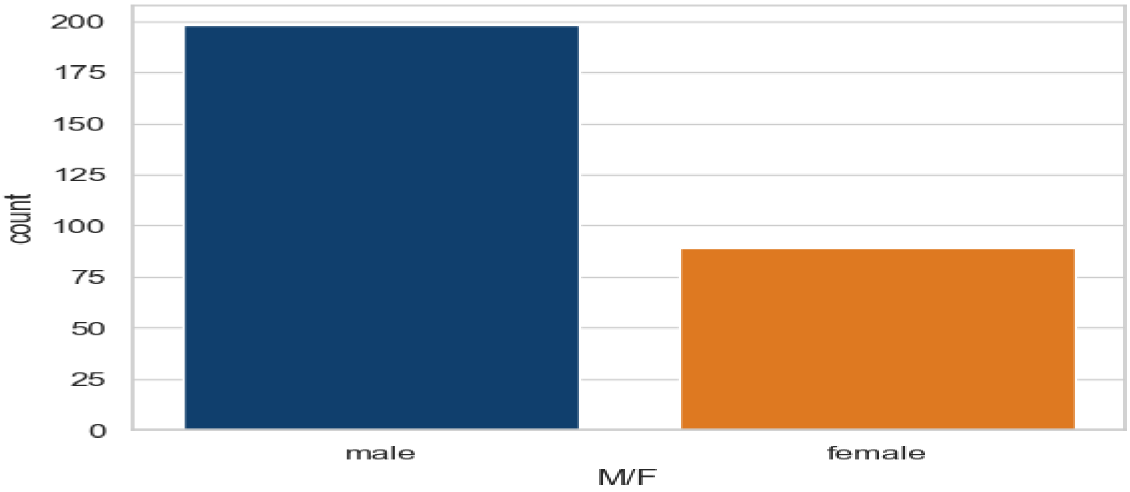
b. FEATURE ENGINEERING

Here in data sex attribute is further classified into 1 as male and 0 as female and also age is just number so make bins of age as decades(10 years)

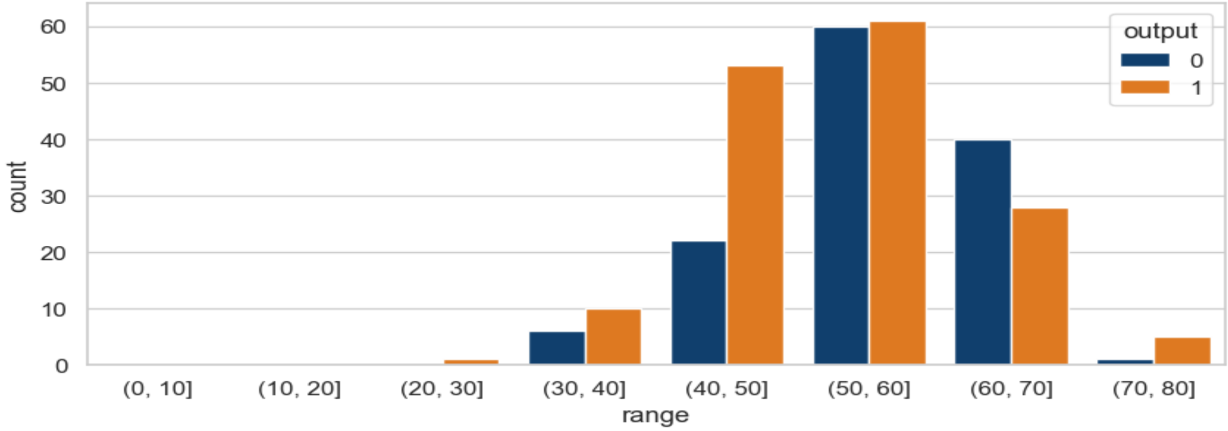


	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output	range	M/F
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1	(60, 70]	male
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1	(30, 40]	male
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1	(40, 50]	female
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1	(50, 60]	male
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1	(50, 60]	female

□ Here we the see the count of target variable 0 or 1 based on sex which is featured as M/F

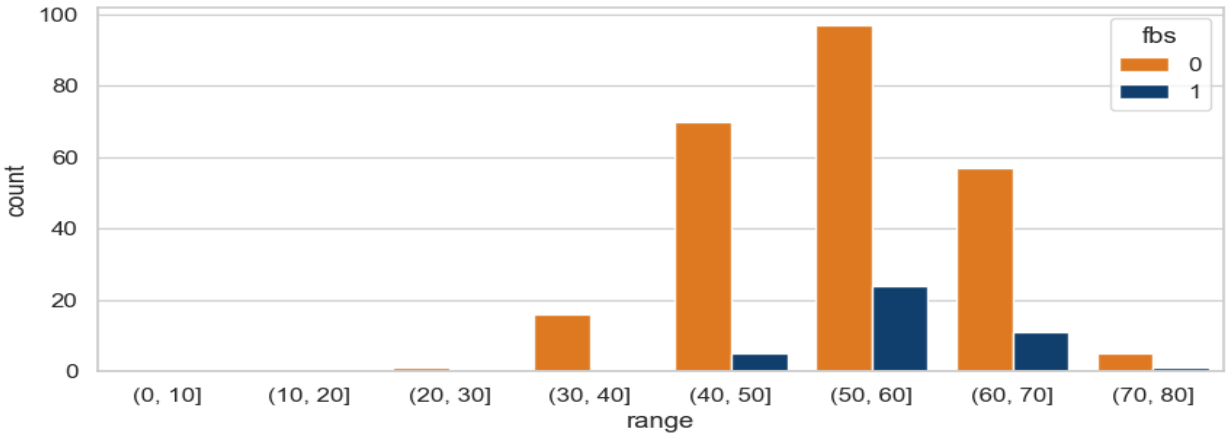


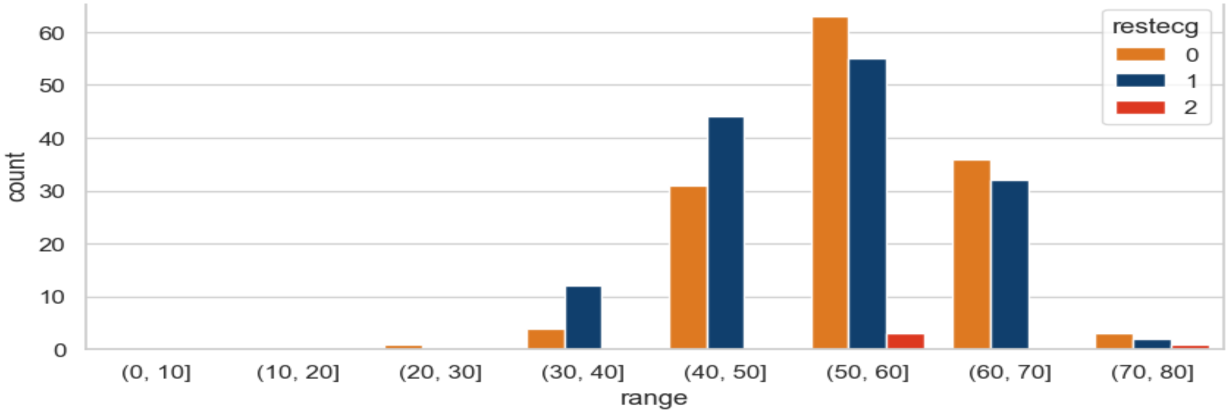
□ Based on this, males are getting higher heart attacks than females.

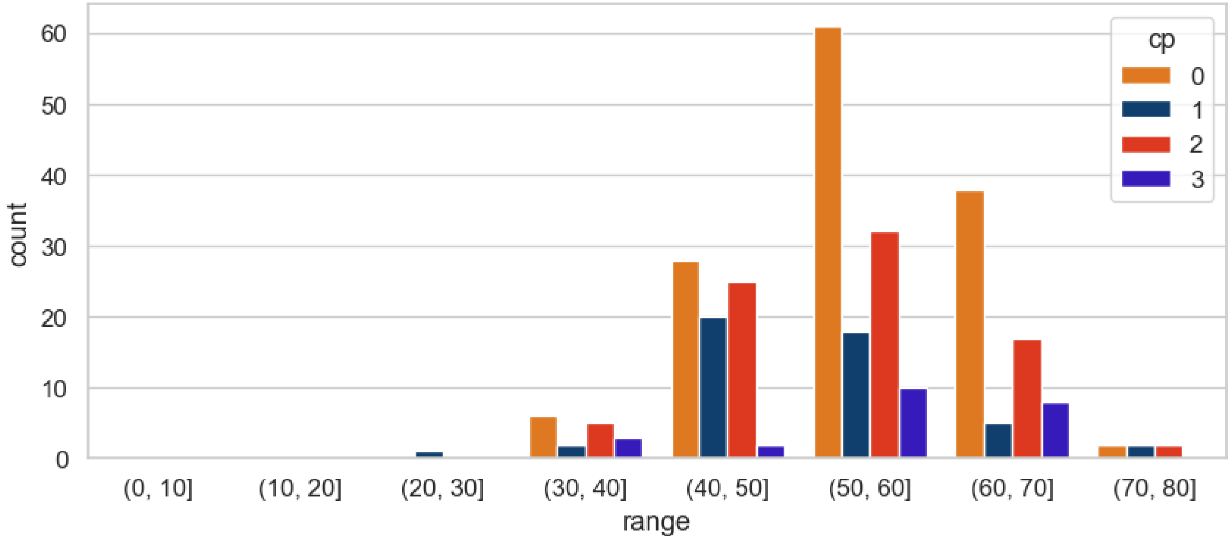


□ Here the plot shows the data is normally distributed over count and range of age.The table on the right side shows the heart attack counts of occur and not not occur with respective gender.

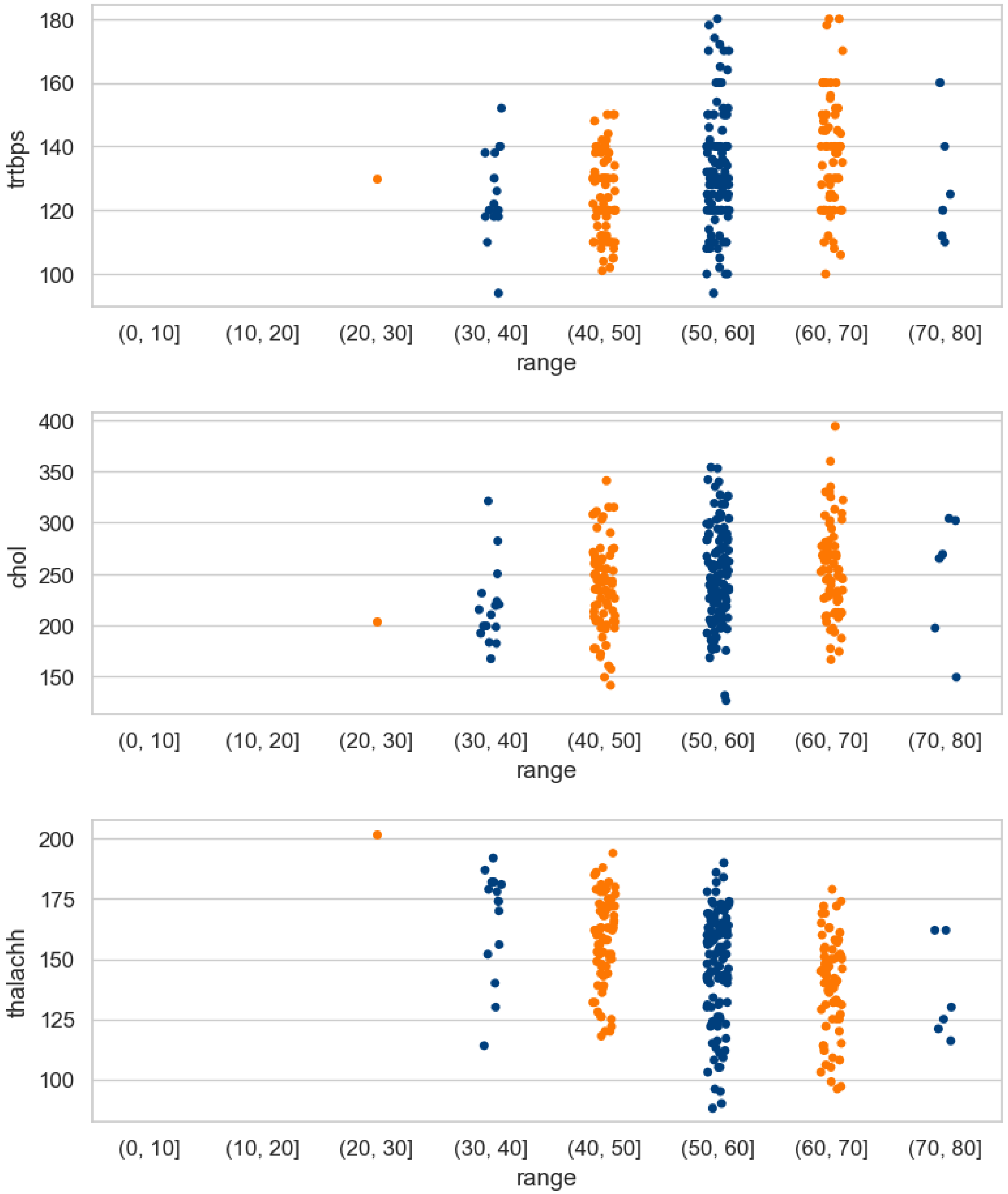
□ These graphs shows,we can see from the above bar chart, people with the age of 50-60 have high chest pain(cp),fasting blood sugar(fbs) and resting electrocardiographic results(restecg)

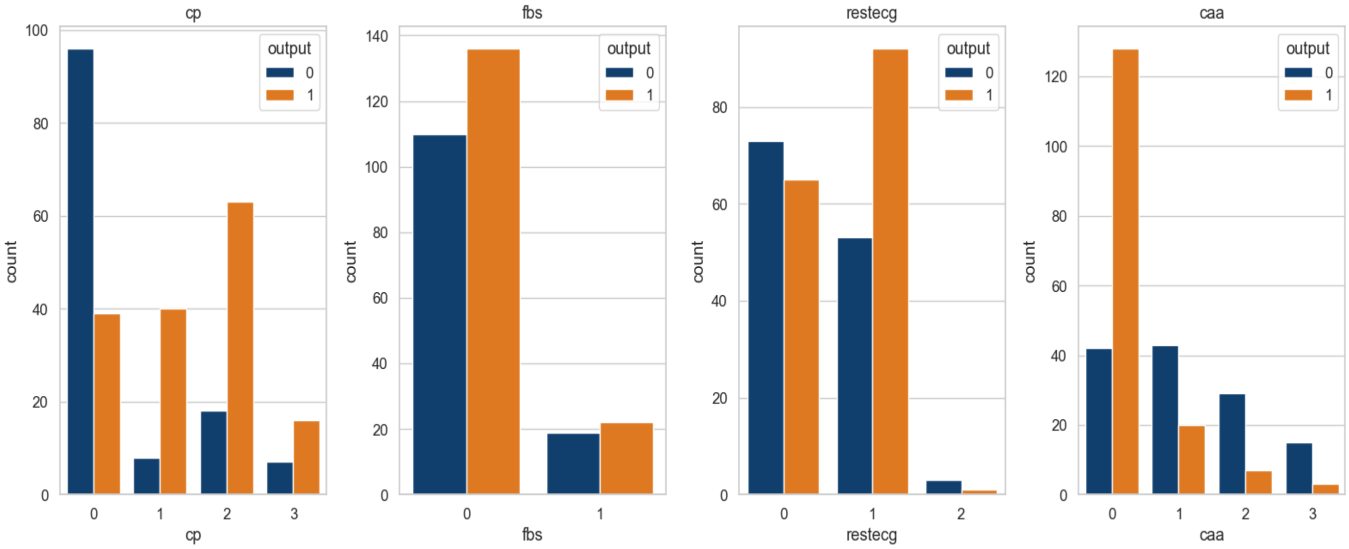






□ These graphs show,as we can see from the above bar chart, people with the age of 30-60 have high thalachh, chol and trtbps.



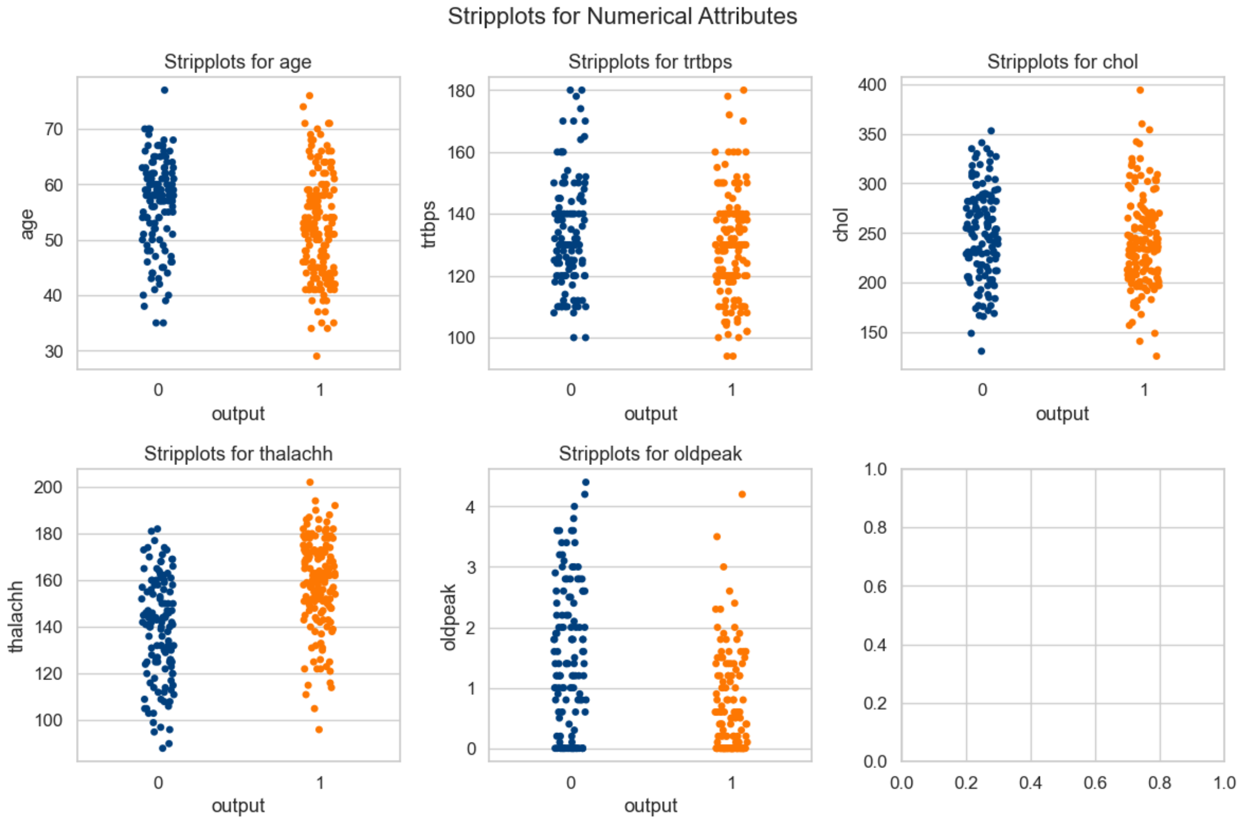


□ People having chest pain(cp) type 2 : atypical angina have high chances of heart attack.

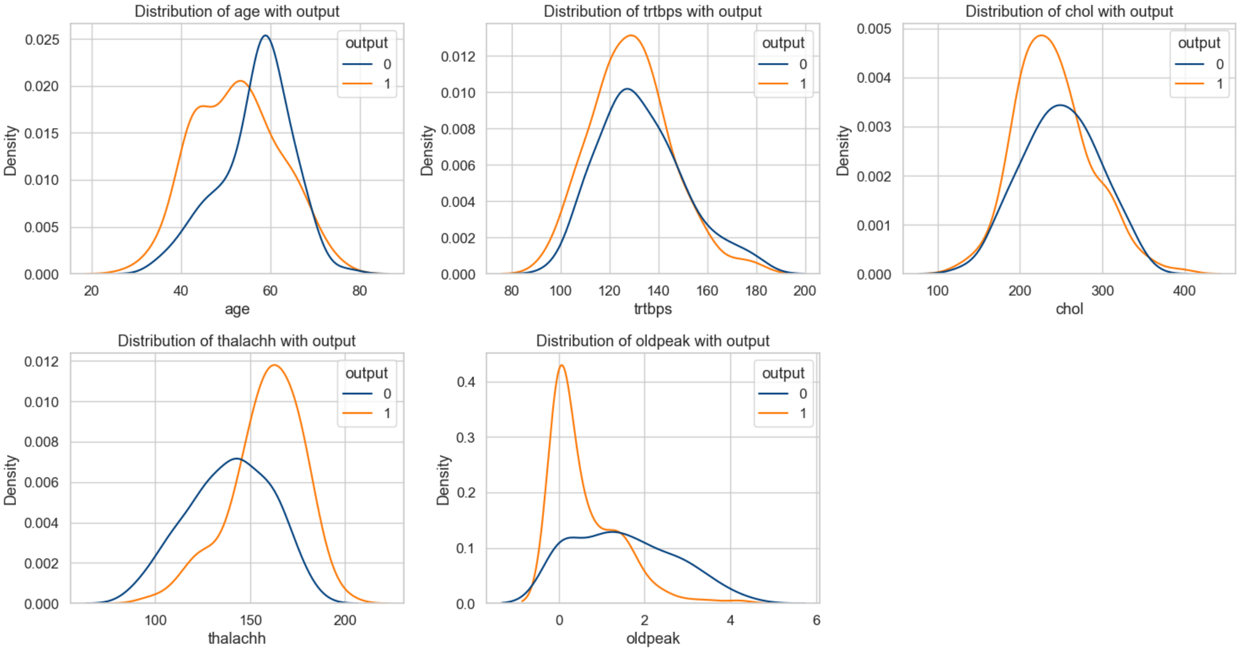
□ People with blood sugar less than 120 mg/dl have chances of heart attack.

□ People with resting electrocardiographic results of value 1 : having ST-T wave abnormality have high chances of heart attack.

□ People with caa type 0 have high chances of heart attack.

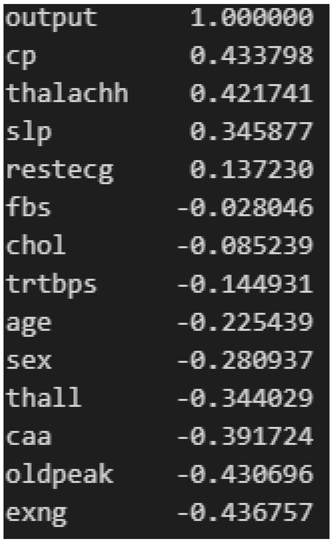


□ These plots are with numerical attributes with output corresponding to heart attack occur or not



□ Distribution of numeric features with the target variable.

□ **CORRELATION :**



```
output      1.000000
cp          0.433798
thalachh    0.421741
slp         0.345877
restecg     0.137230
fbs        -0.028046
chol       -0.085239
trtbps     -0.144931
age        -0.225439
sex        -0.280937
thall      -0.344029
caa        -0.391724
oldpeak    -0.430696
exng       -0.436757
```

3. CLASSIFICATION

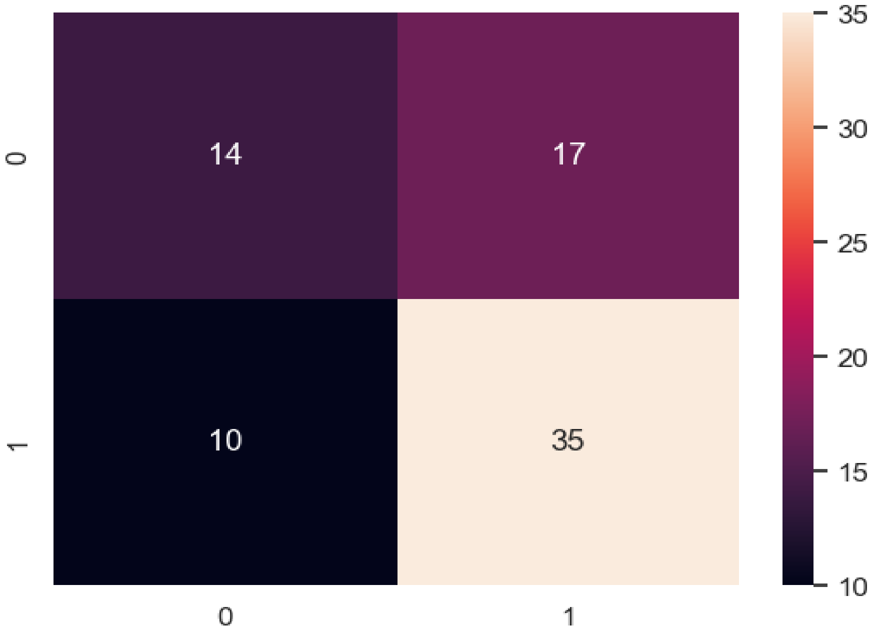
We have used 5 classifiers to check which classifier works best for the prediction of the chance of heart attack. We have divided the data set into testing and training with 75% and 25% respectively.

Classifiers which are used are:

- Support vector Classifier(SVC)
- Random Forest Classifier
- Decision tree(DT)
- K nearest neighbor algorithm(KNN)
- Xgboost

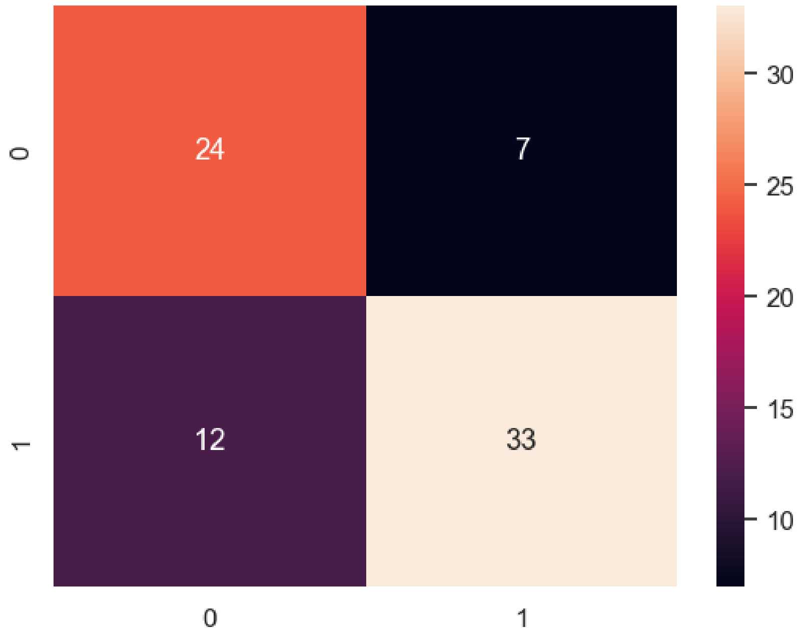
❖ SUPPORT VECTOR CLASSIFIER(SVC)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression analysis. SVM is a powerful and versatile algorithm that can be used for a wide range of applications, including image classification, text classification, and bioinformatics.SVM is a powerful machine learning algorithm for classification and regression analysis. It uses maximum margin classification to find a hyperplane that separates the data into two classes and can handle non-linearly separable data using the kernel trick.



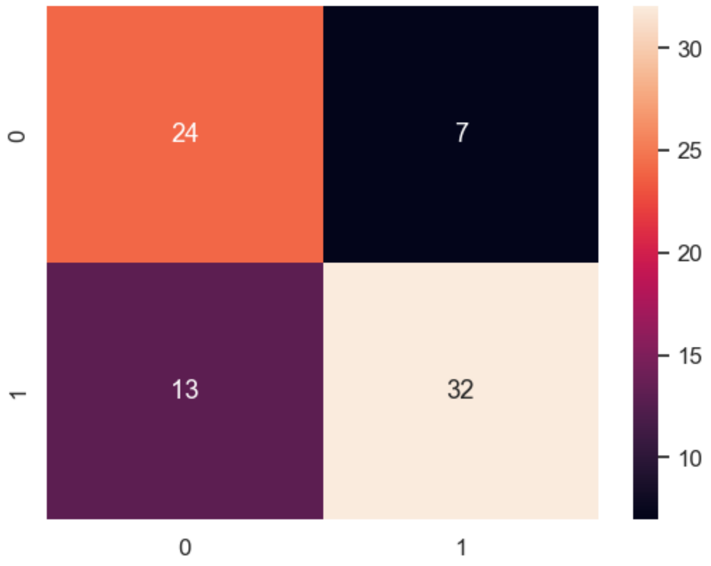
❖ RANDOM FOREST CLASSIFIER

Random Forest Classifier is a machine learning algorithm that combines multiple decision trees to create a more robust and accurate model. It is a type of ensemble learning method that uses bagging to build a forest of decision trees. Each decision tree in the forest is trained on a random subset of the training data and a random subset of the features. The random selection of data and features helps to reduce overfitting and improve the generalization of the model.Random Forest Classifier is a popular algorithm for both classification and regression problems due to its ability to handle high-dimensional data, noisy data, and missing values. It is widely used in applications such as bioinformatics, finance, and remote sensing.



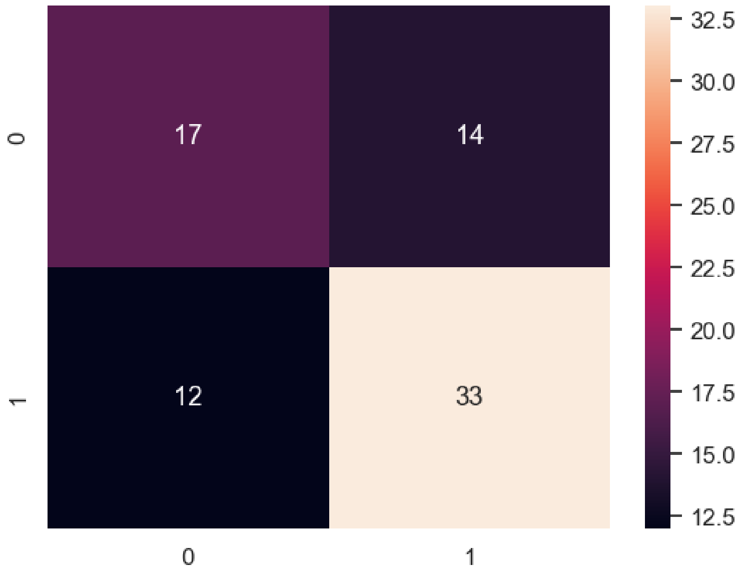
❖ DECISION TREE(DT)

Decision tree is a supervised machine learning algorithm that is used for both classification and regression tasks. It is a type of predictive model that works by recursively partitioning the input data into smaller subsets, based on the values of the input features. Each partitioning creates a node in the decision tree, which represents a decision based on a particular feature value. The tree continues to split until it reaches a stopping condition, such as a maximum depth or a minimum number of samples per leaf. Decision trees can also handle both categorical and numerical data, and can handle missing values by imputing them based on the available data.



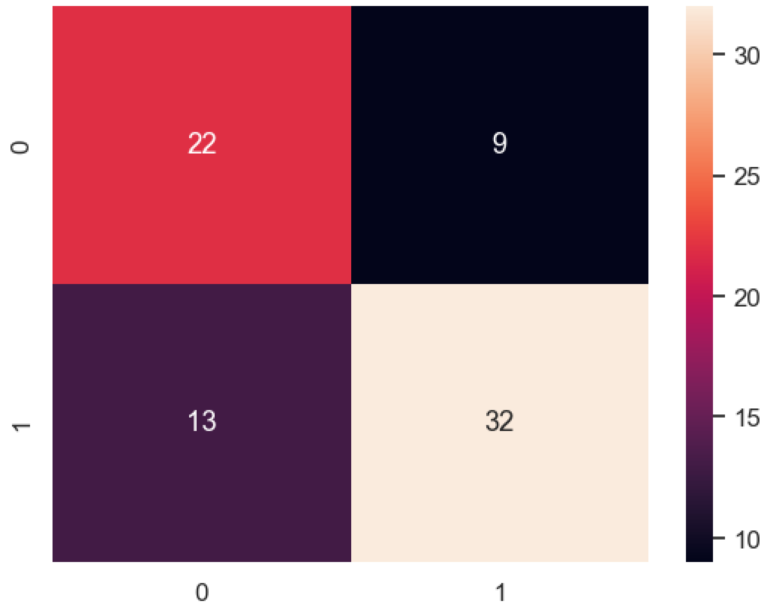
❖ K-NEAREST NEIGHBORS ALGORITHM(KNN)

K nearest neighbor (KNN) algorithm is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the K nearest data points in the training set to a given test data point and using their class labels or numerical values to predict the label or value of the test point. The K in KNN represents the number of nearest neighbors to be considered for prediction.KNN can handle both binary and multi-class classification tasks, as well as regression tasks. KNN can be computationally expensive and may require normalization of the input data to avoid bias towards features with larger magnitudes.



❖ XGBOOST

XGBoost is a popular machine learning algorithm that is used for regression, classification, and ranking problems. It is an implementation of the gradient boosting algorithm that combines multiple weak models to create a more powerful and accurate model. The term XGBoost stands for Extreme Gradient Boosting. It supports parallel processing, can handle sparse data, and provides a range of performance metrics such as AUC, RMSE, and MAE. XGBoost is widely used in various applications such as finance, healthcare, and e-commerce.



RESULTS:

Classifiers	Accuracy	F1 Score
Support vector Classifier(SVC)	0.75	0.8
Random Forest Classifier	0.847222	0.86
Decision tree(DT)	0.763889	0.76
K nearest neighbor algorithm(KNN)	0.694444	0.71
Xgboost	0.819444	0.83

We evaluated the performance of five different classifiers, namely Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, and XGBoost, to predict a target variable. After training and evaluating each classifier, we found that the Random Forest Classifier performed the best.

The Random Forest Classifier had an accuracy of 84.7% and an F1 score of 86%, which was the highest among the five classifiers evaluated. These results indicate that the Random Forest Classifier is a suitable model for predicting the target variable in this study.it is recommended to compare the performance of different classifiers on the same dataset to identify the most appropriate model for a specific problem.

In conclusion, the Random Forest Classifier outperformed the other four classifiers in predicting the target variable, with an accuracy of 84.7% and an F1 score of 86%. These findings suggest that the Random Forest Classifier could be a useful model for predicting similar target variables in future studies.

CONCLUSION

In this study, we aimed to predict the chance of heart attack using a dataset that was subjected to several preprocessing steps. Firstly, we removed influential points to ensure the data was suitable for modeling. Secondly, we conducted exploratory data analysis (EDA) to better understand the distribution and relationships among the variables in the dataset.

The dataset used in this study was found to be balanced, which is an important consideration in machine learning tasks, as imbalanced datasets can result in biased models. We then used five different machine learning models, including Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), Random Forest, and XGBoost, to predict the chance of heart attack.After training and evaluating the models, we found that the Random Forest Classifier performed the best, with an accuracy of 84.7% and an F1 score of 86%. It is important to note that accuracy and F1 score are just two of several evaluation metrics used in machine learning, and the choice of metric depends on the specific problem and application requirements.

In conclusion, this study successfully applied machine learning techniques to predict the chance of heart attack using a balanced dataset, and the best performing model was identified as the Random Forest Classifier. This model can be used to provide early warnings of heart attack risk and enable timely interventions to prevent serious health complications.